Contributions of cochlea-scaled entropy and consonant-vowel boundaries to prediction of speech intelligibility in noise

Fei Chen and Philipos C. Loizou[a)]
Department of Electrical Engineering, University of Texas at Dallas, EC 33, 800 West Campbell Road, Richardson, Texas 75083-0688

(Received 24 June 2011; revised 10 February 2012; accepted 20 February 2012)

Recent evidence suggests that spectral change, as measured by cochlea-scaled entropy (CSE), predicts speech intelligibility better than the information carried by vowels or consonants in sentences. Motivated by this finding, the present study investigates whether intelligibility indices implemented to include segments marked with significant spectral change better predict speech intelligibility in noise than measures that include all phonetic segments paying no attention to vowels/consonants or spectral change. The prediction of two intelligibility measures [normalized covariance measure (NCM), coherence-based speech intelligibility index (CSII)] is investigated using three sentence-segmentation methods: relative root-mean-square (RMS) levels, CSE, and traditional phonetic segmentation of obstruents and sonorants. While the CSE method makes no distinction between spectral changes occurring within vowels/consonants, the RMS-level segmentation method places more emphasis on the vowel-consonant boundaries wherein the spectral change is often most prominent, and perhaps most robust, in the presence of noise. Higher correlation with intelligibility scores was obtained when including sentence segments containing a large number of consonant-vowel boundaries than when including segments with highest entropy or segments based on obstruent/sonorant classification. These data suggest that in the context of intelligibility measures the type of spectral change captured by the measure is important.
[http://dx.doi.org/10.1121/1.3695401]

PACS number(s): 43.71.Gv, 43.71.Es [CYE] Pages: 4104–4113

I. INTRODUCTION

Speech sounds can be separated into two broad categories, i.e., vowels and consonants. The relative contribution of vowels and consonants to speech intelligibility has been controversial. In some respects it is quite challenging to completely isolate the individual contributions of vowels and consonants, since the vowels carry co-articulatory information about consonants at the vowel-consonant boundaries and vowels are inherently longer in duration than consonants. A number of studies based on a noise-replacement paradigm suggested a remarkable advantage of vowels versus consonants for sentence intelligibility. Cole *et al*. (1996) replaced vowel or consonant segments with speech-shaped noise, harmonic complexes or silence in sentences taken from the Texas Instruments/Massachusetts Institute of Technology (TIMIT) corpus (Garofolo *et al*., 1993). Their results showed that the vowel-only sentences (consonants replaced) led to a 2:1 intelligibility advantage over the consonant-only sentences (vowels replaced), regardless of the type of segmental replacement. More specifically, the word recognition rate of vowel-only sentences (consonants substituted with white noise) was 87.4%, around two times that of consonant-only sentences (47.9%) (Cole *et al*., 1996). This intelligibility advantage was maintained even when 10 ms was removed from the onset and offset of the vowel segments. The 2:1 advantage of vowels was later replicated by Kewley-Port *et al*. (2007). Young normal-hearing (NH) and elderly hearing-impaired listeners were presented with sentences at 70 dB and 95 dB sound pressure levels (SPLs), respectively. Participants obtained significantly better performance for the vowel-only sentences, by a ratio of 2:1 across the two groups.

Using the same noise-replacement paradigm, Fogerty and Kewley-Port (2009) further investigated how the perceptual contributions of consonants and vowels were mediated by transitional information present at the consonant-vowel (C-V) boundaries. They defined the speech signal preserved between replacements as a glimpse window. The glimpse windows contained proportional amounts of transitional boundary information either added to consonants or deleted from vowels, yielding two stimulus types, i.e., C + VP and V−VP. The C + VP stimulus preserved the consonant information, included some proportion of the vowel transitions, and replaced the vowel centers with noise. The V−VP stimulus preserved only a proportion of the vowel center information, while replacing the consonants and remaining vowel transitions at the C-V boundary with noise. Fogerty and Kewley-Port (2009) found that the identification accuracy increased linearly for the C + VP stimuli in proportion to the amount of vowel transitions added. The intelligibility of the vowel-only (i.e., V−VP) stimuli was unaffected when less than 30% of the CV transitions was replaced with noise. This was interpreted to indicate that CV transitions provided information redundant with the information present in the vowel centers, an outcome consistent with that reported by Strange *et al*. (1983) with CVCs.

a)Author to whom correspondence should be addressed. Electronic mail: loizou@utdallas.edu

Grounded on the well-known fact that perceptual systems respond primarily to change, Stilp and Kluender (2010) recently suggested that cochlea-scaled entropy (CSE), not vowels, consonants or segment duration, best predicts speech intelligibility. They measured cochlea-scaled entropy in TIMIT sentences and replaced portions of the sentences having high, medium, or low entropy with equal-level noise. Replacing low-entropy segments yielded relatively small impact on intelligibility while replacing high-entropy segments significantly reduced sentence intelligibility. A remarkably robust correlation was found with cochlea-scaled entropy predicting listeners' intelligibility scores. Stilp and Kluender (2010) also reported that the duration of the signal replaced and proportion of consonants/vowels replaced were not significant predictors of intelligibility, and thus did not account for the strong relationship between CSE and sentence intelligibility.

The outcomes of the above studies are important in the design of intelligibility measures, particularly the ones that rely on short-term (20–30 ms) processing (Rhebergen and Versfeld, 2005; Taal *et al.*, 2010; Kates and Arehart, 2005; Ma *et al.*, 2009). Such measures typically place equal emphasis on all segments processed, paying the same attention to transitional segments marked with significant spectral change (e.g., vowel-consonant boundaries) and to steady-state (or quasi steady-state) segments (e.g., vowel centers). This is, however, contrary to existing speech perception literature pointing to differences in the contributions of vowels vs consonants (e.g., Kewley-Port *et al.*, 2007) and differences between low and high-entropy segments (Stilp and Kluender, 2010) on speech recognition. If vowels do indeed carry more information than consonants, that would suggest the development of intelligibility measures that place more emphasis on the vocalic segments rather than the consonant segments. Such an emphasis could be implemented by somehow placing a larger weight on those information-bearing segments. Similarly, if high-entropy segments carry perceptually more information than low-entropy segments, then one can devise a measure that applies greater weight on high-entropy segments than low-entropy segments.

The aim of the present study is to identify the segments that carry most of the information in sentences and as such should be included (or emphasized) in the computation of intelligibility indices. The underlying hypothesis is that including only these information-bearing segments in the computation of intelligibility indices ought to improve the correlation with human listener's intelligibility scores relative to the scenario where all segments are included. Unlike previous studies (e.g., Kewley-Port *et al.*, 2007; Stilp and Kluender, 2010) that replaced the segments of interest with equal-level noise and assessed their importance with listening experiments, the present study evaluates indirectly the perceptual importance of these segments in the context of intelligibility measures with the main goal of improving the prediction power of existing intelligibility measures. Clearly, the method used for segmenting sentences (whether phonetically or not) into different units will affect the predictive power of the intelligibility index. The present study examines the intelligibility prediction performance of two intelligibility measures implemented using three different sentence segmentation methods: one based on relative root-mean-square (RMS) levels, one based on cochlea-scaled entropy and one based on the traditional phonetic segmentation of obstruents and sonorants. Sentence segmentation based on explicit vowel/consonant boundaries is not pursued here since that it is extremely challenging to implement in practice even with using the most sophisticated phoneme detection algorithms. In contrast, all three segmentation methods examined in the present study can be applied via an algorithm to arbitrary speech stimuli without expert knowledge of acoustics/phonetics. By examining and analyzing different segmentation strategies, we can assess indirectly the perceptual contributions of various types of phonetic segments to sentence intelligibility prediction.

II. SPEECH INTELLIGIBILITY DATA

The speech intelligibility data was taken from the intelligibility evaluation of noise-corrupted speech processed through eight different noise-suppression algorithms by a total of 40 NH listeners (Hu and Loizou, 2007). IEEE sentences (IEEE, 1969) were used as test material, and all sentences were produced by a male talker. The sentences were originally sampled at 25 kHz and down-sampled to 8 kHz. The masker signals were taken from the AURORA database (Hirsch and Pearce, 2000) and included the following real-world recordings from different places: babble, car, street, and train. The maskers were added to the speech signals at signal-to-noise ratio (SNR) levels of 0 and 5 dB. The processed speech sentence files, along with the noisy speech files, were presented monaurally to the listeners in a double-walled sound-proof booth (Acoustic Systems, Inc.) via Sennheiser's HD 250 Linear II circumaural headphones at a comfortable listening level. Two IEEE sentence lists (ten sentences per list) were used for each condition, and none of the sentence lists were repeated. The intelligibility scores were obtained from NH listeners in a total of 72 conditions ($=4$ maskers $\times$ 2 SNR levels $\times$ 8 algorithms $+$ 4 maskers $\times$ 2 noisy references). The percentage intelligibility score for each condition was calculated by dividing the number of words correctly identified by the total number of words in a particular testing condition. More details about the noise-suppression algorithms and the procedure used to collect the intelligibility data can be found in Hu and Loizou (2007). The sentences used in that study were produced by a male talker, but we do not expect talker-specific factors to influence the prediction power of the intelligibility measures considered in the present study. This is so because none of the measures considered here extracts F0-dependent features from the signal.

III. SPEECH INTELLIGIBILITY MEASURES

Present intelligibility indices employ primarily either temporal-envelope or spectral-envelope information to compute the index. For the temporal-envelope based measure, we examined the intelligibility prediction performance of the normalized covariance measure (NCM), which is classified as a speech-transmission index (STI)[1] based measure (see review in Goldsworthy and Greenberg, 2004). For the spectral-envelope based measure, we investigated the coherence-based

speech intelligibility index (CSII) measure (Kates and Arehart, 2005).

The NCM index is similar to the speech-transmission index (Steeneken and Houtgast, 1980) in that it computes a weighted sum of transmission index (TI) values determined from the envelopes of the probe (input) and response (output) signals in each frequency band (Goldsworthy and Greenberg, 2004). Unlike the traditional STI measure, however, which quantifies the change in modulation depth between the probe and response envelopes using the modulation transfer function, the NCM index is based on the covariance between the probe and response envelope signals computed in each band. In its original implementation, the NCM index makes use of the envelopes extracted for the whole utterance to compute the TI value of each band (TI values are subsequently converted to an apparent SNR and mapped to the NCM index taking values between 0 and 1). In the present study, we modified the NCM index as follows to account for different segmentation methods, i.e., to account for a select set of segments entering its computation. Using the probe signal, the time instances of the segments of interest (e.g., obstruent and sonorant segments) are first determined according to the segmentation methods described below. The envelopes falling within each of the selected segments were then concatenated into one composite envelope of each frequency band. This was done for both the probe and response stimuli. Finally, the corresponding composite (concatenated) probe and response envelopes were used to compute the TI values for each band and subsequently the NCM index. We believe that when using the new composite envelope, formed by concatenating the various segments together, the modified NCM index computes a perceptually more relevant (apparent) SNR in each band since only information-bearing segments are included in its computation.

The speech intelligibility index (SII) (ANSI, 1997) is based on the principle that the intelligibility of speech depends on the proportion of spectral information that is audible to the listener and is computed by dividing the spectrum into 20 bands (contributing equally to intelligibility) and estimating the weighted average of the signal-to-noise ratios in each band (Kryter, 1962a,b; Pavlovic, 1987; ANSI, 1997). The modified coherence-based SII index (CSII) (Kates and Arehart, 2005) uses the base form of the SII procedure, but with the signal-to-noise ratio term replaced by the signal-to-distortion ratio, which was computed using the coherence function between the input and processed signals. The CSII measures have been used extensively to assess subjective speech quality (Arehart *et al.,* 2007) and speech distortions introduced by hearing aids (Kates, 1992; Kates and Arehart, 2005). These measures have also been shown to yield high correlations with the intelligibility of vocoded speech (Chen and Loizou, 2011a), vocoded and wideband (non-vocoded) Mandarin Chinese (Chen and Loizou, 2011b), and noise-masked speech processed by noise reduction algorithms (Ma *et al.,* 2009). More details regarding the definition and implementation of the NCM and CSII measures can be found in Ma *et al.* (2009).

This study assesses the intelligibility prediction performance of the NCM and CSII measures implemented using three different sentence segmentation methods. All combinations of segmentations were incorporated in each measure irrespective of the measure being temporal-envelope based (NCM) or spectral-envelope based (CSII). The CSII measure, for instance, was implemented using a temporal RMS-level based segmentation scheme (Kates and Arehart, 2005) as well as the spectrally-based CSE segmentation scheme. A detailed description of the three segmentation methods examined is given next.

A. Scaled-entropy based segmentation

To compute the cochlea-scaled spectral entropy (Stilp and Kluender, 2010), the sentence is first normalized according to its RMS intensity, and then divided into 16 ms segments. Segments are first bandpass filtered into M bands using ro-ex filters (Patterson *et al.,* 1982). The ro-ex filters capture the non-linear weighting and frequency distribution along the cochlea. In this study, $M = 16$ filters are used, spaced one equivalent rectangular bandwidth (ERB) apart, spanning 300 to 3400 Hz.[2] Euclidean distances between adjacent 16 ms segments are calculated across the M filter-output levels. Distances are then summed in boxcars of five successive segments (80 ms in duration). Cumulative Euclidean distances within a boxcar are taken as the measures of spectral entropy. More details on the computation of the cochlea-scaled entropy can be found in Stilp and Kluender (2010).

Figure 1(b) illustrates the cochlea-scaled entropy computation for one sentence. The sentence is segmented into two regions, i.e., low-entropy (L-entropy) and high-entropy (H-entropy) regions according to an entropy threshold E_{thr}. The entropy threshold E_{thr} is determined according to a proportion coefficient p (given in percent) such that p percent of all entropies within the utterance are smaller than E_{thr}. Hence, assuming that the entropy values are sorted in ascending order, a value of $p = 0.6$ would suggest that 60% of the entropy values (within the utterance) are smaller than E_{thr}. With the above-prescribed entropy threshold, the low and high-entropy regions are defined to include those segments whose entropies are either smaller or larger than E_{thr}, respectively, as shown in Fig. 1(b). In the present study, the proportion coefficient varied from $p = 0.1$ to $p = 0.9$ in steps of 0.1.

B. Relative RMS-level based segmentation

The relative-RMS-level-based segmentation is implemented by dividing speech into short-term (16 ms in this study) segments and classifying each segment into one of three regions according to its relative RMS intensity (Kates and Arehart, 2005). The high-level (H-level) region consists of segments at or above the overall RMS level of the whole utterance. The mid-level (M-level) region consists of segments ranging from the overall RMS level to 10 dB below (i.e., RMS-10 dB), and the low-level (L-level) region consists of segments ranging from RMS-10 dB to RMS-30 dB. We adopt the same threshold levels (i.e., 0, −10 and −30 dB) as proposed by Kates and Arehart (2005). Figure 1(c) shows an example sentence segmented into H-, M- and L-levels based on the above RMS threshold levels. For the

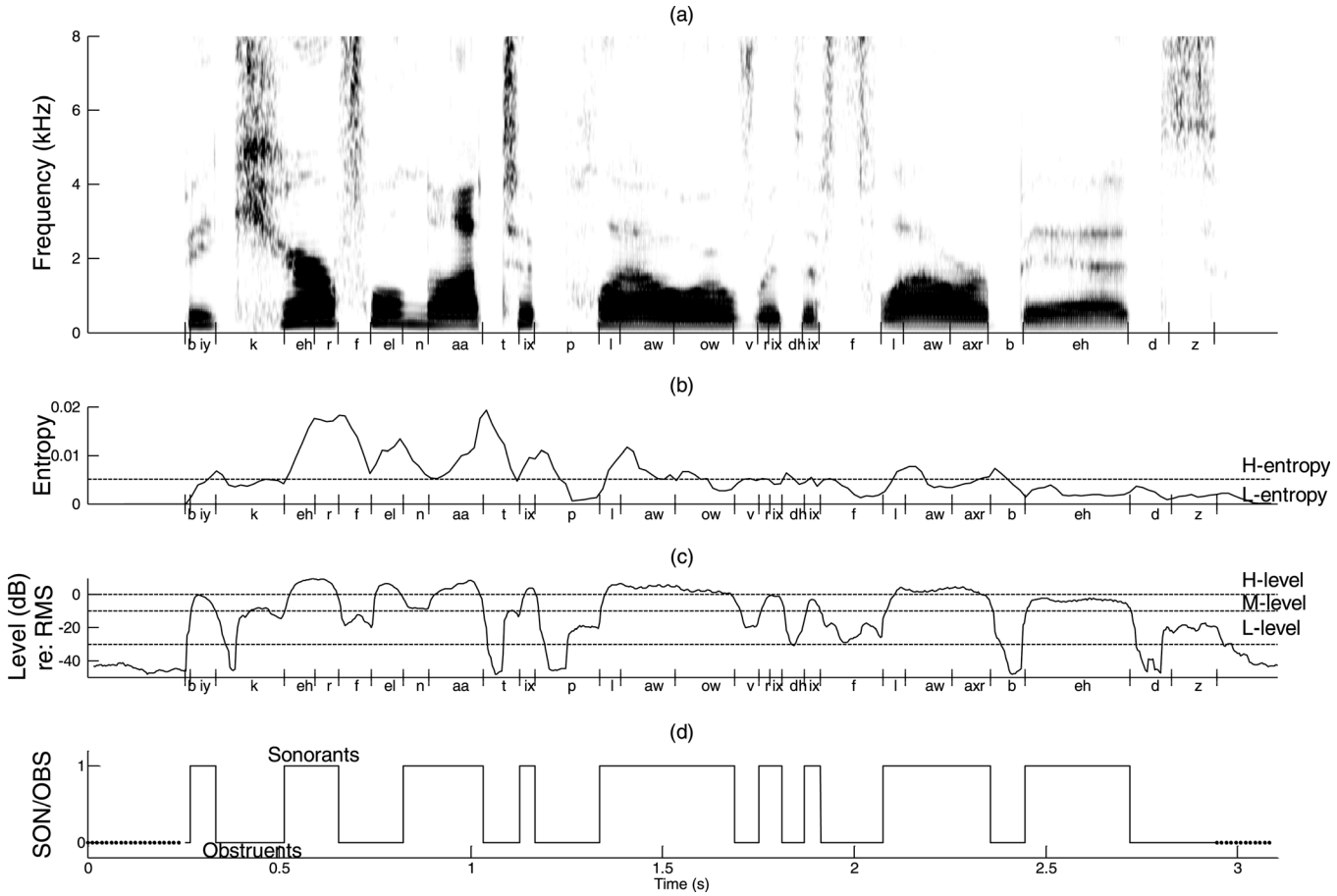

FIG. 1. (a) The spectrogram of the TIMIT sentence "Be careful not to plow over the flower beds." (b) The segmentation of the sentence based on scaled-entropy, (c) RMS levels, and (d) sonorants/obstruents classification. The proportion coefficient in (b) was set to 60%. The sonorants and obstruents in (d) are indicated as levels of 1 and 0, respectively, and the dotted line indicates the silent regions at the beginning and end of the sentence.

most part, H-level segments include vowels and semivowels, M-level segments include consonants and vowel-consonant transitions, and L-level segments include primarily weak consonants (more on this later).

C. Sonorant/obstruent based segmentation

Speech sounds can generally be divided into two broad classes, i.e., obstruents and sonorants based on the presence/absence of periodicity (Maddieson, 1984). Obstruents (e.g., stops, fricatives and affricates) are typically voiceless (aperiodic), while sonorants (e.g., vowels) are typically voiced (periodic). Figure 1(d) shows an example of sentence segmentation into sonorant and obstruent regions. In the present study, we will be using the hand-labeled sonorant/obstruent segmentations of the IEEE sentence corpus available from Loizou (2007).

IV. RESULTS

The average intelligibility scores obtained by NH listeners in Sec. II were subjected to correlation analysis with the corresponding values obtained by the NCM and CSII measures implemented using the above sentence-segmentation methods. More specifically, correlation analysis was performed between the mean (across all subjects) intelligibility scores obtained in each of the 72 testing conditions (Sec. II) and the corresponding mean (computed across the 20 sentences used in each condition) intelligibility index values obtained in each condition. The Pearson's correlation coefficient (r) was used to assess the performance of the intelligibility measures to predict intelligibility scores. Table I shows the correlation coefficients r between speech recognition scores and NCM/CSII measures implemented with the three segmentation methods described in Sec. III. For comparative purposes, Table I also includes the correlation coefficients computed using all segments of the sentences, i.e., when no differentiation is made between low- and high-entropy segments, or among low-, middle- and high-RMS level segments, or between obstruent and sonorant segments. These correlation coefficients served as control and were $r = 0.80$ and $r = 0.82$ for the NCM and CSII measures, respectively.

For the correlation coefficients obtained using the scaled entropy-based segmentation in Table I, the proportion coefficient was set to $p = 60\%$ (correlations with other p values are provided later in Fig. 2). As can be seen, compared with the CSII correlation obtained using all segments ($r = 0.82$), no improvement was noted when using the high-entropy segments to predict intelligibility. When incorporated into the implementation of the NCM index, however, the entropy segmentation slightly improved the resulting correlation, i.e., $r = 0.82$ using high-entropy segments versus $r = 0.80$ using all segments.

TABLE I. Correlation coefficients (r) between listeners' sentence recognition scores and intelligibility index values computed with three different segmentation strategies. The scaled entropy was computed with proportion coefficient $p = 60\%$.

		Scaled entropy		RMS-level			Sonorant/Obstruent	
Intelligibility measure	All segments	H-entropy	L-entropy	H-level	M-level	L-level	Sonorants	Obstruents
NCM	0.80	0.82	0.78	0.83	0.89	0.77	0.84	0.64
CSII	0.82	0.82	0.82	0.85	0.91	0.86	0.85	0.69

As reported in our previous study (Ma *et al.*, 2009), the M-level based CSII measure performed the best ($r = 0.91$) compared to the other two RMS-level segmentations, i.e., $r = 0.91$ (M-level) vs 0.85 (H-level) and 0.86 (L-level) in Table I. As can be seen, the NCM indices when implemented using M-level segmentation also correlated highly with intelligibility scores ($r = 0.89$). Ma *et al.* (2009) proposed several signal-dependent band weighting functions (BWFs) for predicting the intelligibility of speech corrupted by fluctuating maskers. The highest correlation obtained with the NCM index was $r = 0.89$ when signal-dependent BWFs were used for predicting the same dataset of intelligibility scores (Ma *et al.*, 2009). In this regard, the above finding shows that the M-level segmentation can be used as an alternative and simpler method for improving the intelligibility prediction performance of the NCM index. This benefit (i.e., using M-level RMS segmentation instead of signal-dependent BWFs to predict sentence intelligibility) has also been noted in other studies, involving datasets of vocoded English (Chen and Loizou, 2011a) and vocoded and wideband Mandarin Chinese (Chen and Loizou, 2011b).

It is evident from Table I that the sonorant and obstruent segments had a differential impact on the correlation with sentence scores. The correlation coefficients obtained by both measures using only sonorant segments were higher than those obtained when using all segments. More precisely, the resulting correlations were $r = 0.84$ vs 0.80 for the NCM index and $r = 0.85$ vs 0.82 for the CSII measure. Use of obstruent segments alone led to a much lower correlation compared with that using all segments, i.e., $r = 0.64$ vs 0.80 and $r = 0.69$ vs 0.82 for the NCM and CSII measures, respectively.

Tables II and III show the statistical comparison between the correlation coefficients reported in Table I. Statistical analysis was performed as per Steiger (1980). When compared to the standard normal curve rejection points of ± 1.96, the correlation coefficients of the NCM and CSII indices implemented with the M-RMS-level segmentation were found to be significantly ($p < 0.05$) higher than those obtained with other types of segmentation methods.

A. Scaled-entropy analysis

To investigate the influence of the proportion coefficient p used in entropy-based segmentation on intelligibility prediction, we examined the correlation of the CSII and NCM indices with sentence intelligibility for different values of the proportion coefficient p ranging from 10% to 90%. The resulting correlations are plotted in Fig. 2. The intelligibility prediction of the L-entropy CSII measure [Fig. 2(a)] improved with increasing values of the proportion coefficient p, i.e., as more L-entropy segments were added. Correlation improved, for instance, from $r = 0.10$ to 0.82 as p varied from 10% to 60%, respectively. This suggests that in order to achieve a fairly good intelligibility prediction, segments with at least 60% lowest entropy need to be included in the CSII computation. The correlation pattern with H-entropy segments was found to be flat. At one end, the correlation was nearly the same as the control correlation (i.e., $r = 0.81$

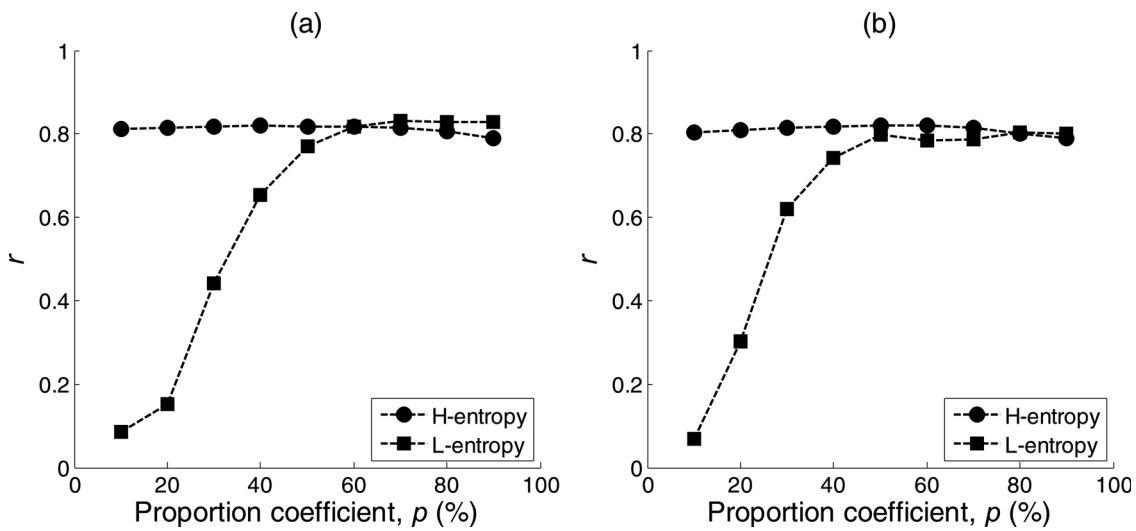

FIG. 2. Correlation coefficients (r) between listeners' sentence recognition scores and (a) CSII values and (b) NCM values computed using L-entropy and H-entropy segmentation methods for varying values of the proportion coefficient p.

TABLE II. Statistical comparison (based on $\alpha = 0.05$) between the correlation coefficients of the NCM index (with sentence intelligibility scores) implemented with various types of segmentation (as reported in Table I). Asterisk indicates that the difference in correlation is statistically significant ($p < 0.05$).

NCM	All segments	H-entropy	L-entropy	H-RMS-level	M-RMS-level	L-RMS-level	Sonorants	Obstruents
All segments	–							
H-entropy	1.46	–						
L-entropy	0.61	1.22	–					
H-RMS-level	1.32	0.73	1.26	–				
M-RMS-level	4.97*	4.12*	4.44*	3.50*	–			
L-RMS-level	0.78	1.33	0.40	1.44	4.45*	–		
Sonorants	3.00*	1.83	1.82	0.55	4.04*	1.87	–	
Obstruents	3.79*	3.85*	3.72*	3.71*	6.34*	4.22*	4.40*	–

and 0.82 for CSII implemented with H-entropy and all segments, respectively), while at the other end the correlation slightly dropped to $r = 0.79$ when the proportion coefficient p reached 90% (i.e., when segments with 10% largest entropies were utilized in the computation of the CSII measure). Reducing p to 60% only slightly increased the correlation coefficient to $r = 0.82$ and further extending to include more H-entropy segments did not seem to improve the correlation of the CSII measure. This finding indicates that a small amount (i.e., 10%) of high entropy segments carries sufficient information to yield high intelligibility prediction ($r = 0.79$). The same correlation pattern was also observed for the NCM index, as shown in Fig. 2(b). Overall, the data in Fig. 2 are consistent with the findings of Stilp and Kluender (2010): high entropy segments contribute more to intelligibility than low entropy segments. The cutoff point, however, beyond which no benefit is observed seems to be around $p = 60\%$. That is, correlation is about the same regardless of whether the 40% highest entropy segments are used in the CSII computation or the 60% lowest entropy segments are used.

B. RMS segmental level analysis

As mentioned earlier, H-level segments contain primarily vowels, M-level segments contain a subset of the consonants along with vowel-consonant transitions, and L-level segments contain weak consonants. This can be easily inferred from the RMS plot shown in Fig. 1. The accurate distribution, however, of the various sound classes (e.g., vowels, consonants) in the three levels is lacking and was not clearly quantified in the literature. For that reason, we performed a detailed analysis of the distribution of phonemes contained in the H-, M- and L- level segments using the TIMIT corpus (Garofolo *et al.*, 1993). The TIMIT corpus was used because of lack of accurate phonetic transcriptions with the IEEE corpus. A total of 40 TIMIT sentences were used in the analysis, and Table IV tabulates the classification of TIMIT symbols into the various sounds. The 40 TIMIT sentences were extracted from the Southern dialect region (DR5). Each sentence was produced by a different talker, and overall, 20 male and 20 female talkers were used. Similar to the study by Kewley-Port *et al.* (2007), consonant strings were treated as a single unit, as were vowel strings. Therefore, a C-V boundary only occurred between consonant and vowel units, such as a vowel followed by a consonant or vice versa (note that the C-V boundary is defined here to be bi-directional, i.e., consonant transition into a vowel and vice versa). The 40 sentences contained an average of 7.8 words/sentence (range: 6–10 words), 37.5 phonemes/sentence (12.0 vowels, 25.5 consonants), and 22.7 C-V boundaries/sentence. The phoneme distributions of the selected TIMIT sentences were consistent with those reported by Kewley-Port *et al.* (2007).

Table V shows the distribution of vowels and consonants present in the three types of RMS-level segments. This distribution was computed in terms of relative duration expressed in percentage of vowels or consonants present in a specific type of RMS-level segment. The non-phonetic segments in Table V include pauses and epenthetic silence contained in TIMIT sentences. As shown in Table V, H-level segments are dominated by vowels (79.5%), while L-level segments are dominated by consonants (84.5%). The duration of vowels and consonants was nearly the same (46.2% vs 53.7%) in M-level segments. Table V also lists the percent durations of the consonants when further sub-classified into semivowels, nasals, stops, fricatives and affricates. Semivowels are the

TABLE III. Statistical comparison (based on $\alpha = 0.05$) between the correlation coefficients of the CSII measure (with sentence intelligibility scores) implemented with various types of segmentation (as reported in Table I). Asterisk indicates that the difference in correlation is statistically significant ($p < 0.05$).

CSII	All segments	H-entropy	L-entropy	H-RMS-level	M-RMS-level	L-RMS-level	Sonorants	Obstruents
All segments	–							
H-entropy	0.02	–						
L-entropy	0.01	0.02	–					
H-RMS-level	1.49	0.91	1.87	–				
M-RMS-level	4.84*	4.11*	5.03*	3.88*	–			
L-RMS-level	1.40	1.27	1.47	0.38	5.03*	–		
Sonorants	1.87	0.89	2.83*	0.15	4.54*	0.47	–	
Obstruents	3.00*	2.66*	3.19*	3.76*	6.12*	3.90*	4.02*	–

TABLE IV. Classification of TIMIT symbols.

Vowels	iy, ih, eh, ae, aa, er, ax, ah, ao, uw, uh, ow, ay, oy, aw, ey, ux, ix, axr, ax-h, axr, ax-h
Consonants	
Semivowels	w, l, r, y, hh, hv, el
Nasals	m, em, n, en, nx, ng, eng
Stops	b, d, dx, g, p, t, k, bcl, dcl, gcl, pcl, tcl, kcl
Fricatives	v, dh, z, zh, f, th, s, sh
Affricates	jh, ch

dominant (i.e., 14.9%) consonants present in H-level segments, while the stops, fricatives and affricates when combined make up only a small portion (i.e., 5.2%) of the H-level segments. As sonorant segments consist of vowels, semivowels and nasals, the H-level segmentation captures mainly sonorant segments (i.e., 94.8%). Thus, it is not surprising that the correlation coefficients obtained using H-level segments are nearly equal to that obtained using sonorant segments (i.e., $r = 0.83$ and 0.84 for the NCM index, and $r = 0.85$ for the CSII measure in Table I).

As expected, more consonants are present in M-level segments than in H-level segments. The duration (in percentage) of stops and fricatives is increased to 9.3% and 19.4%, respectively. M-level segments also contain fewer vowels (46.2%). Based on the threshold value used for classifying M-level segments, more vowel-consonant transitions are to be expected to be present in M-level segments. To further quantify this, Table VI shows the distribution of vowel-consonant boundaries present in the three RMS levels. The percentages shown in Table VI were computed by dividing the number of C-V boundaries contained in a specific segment (e.g., M-level) by the total number of C-V boundaries (numbers were averaged over 40 TIMIT sentences). As shown in Table VI, more than half (i.e., 53.6%) of the total number of C-V boundaries present in sentences are contained in M-level segments. In comparison, the numbers of C-V boundaries present in H-level and L-level segments are much less, i.e., 18.9% (H-level) and 27.5% (L-level) vs 53.6% (M-level). Hence, in brief, M-level segments capture the majority of the C-V transitions present in sentences.

Finally, Table VII shows the H-, M- and L-level segment durations of the IEEE sentences used in this study when no differentiation is made between vowels and consonants. The L-level segments occupy the largest portion of the sentences (39.1%). The M-level and H-level segments were nearly equal in duration (28.9% vs 29.1%). Yet, the corresponding correlations were markedly different (e.g., $r = 0.89$ vs $r = 0.83$ for the NCM indices implemented with M-level and H-level segments, respectively), suggesting that the overall duration of the segments included in the computation of the intelligibility measures did not influence the prediction of sentence intelligibility. For comparison, Table VII also shows the obstruents/sonorants duration distribution. As expected, the duration of sonorant segments was longer than that of obstruent segments, i.e., 57.2% vs 36.5%. The sentence intelligibility prediction was better when sonorant segments were included, but this was also better than that obtained when *all* segments were included. This provides additional confirmation that the overall duration of the segments included in the computation of the intelligibility measures did not influence the prediction power of the measures.

TABLE VI. Distribution of C-V boundaries in terms of percentage of C-V boundaries present in each type of scaled-entropy and RMS-level segments. The scaled entropy was computed with proportion coefficient $p = 60\%$.

	Scaled entropy		RMS-level		
	H-entropy	L-entropy	H-level	M-level	L-level
C-V boundaries	49.8%	50.2%	18.9%	53.6%	27.5%

V. DISCUSSION AND CONCLUSIONS

A. Contributions of vowels and consonants to sentence intelligibility prediction

Based on the above analysis, we know that the H-level segments are dominated by vowels while the L-level segments are dominated by (weak) consonants (Table V). With the exception of one condition with the CSII measure, better intelligibility prediction was obtained with H-level segments ($r = 0.83$, NCM index) than with L-level segments ($r = 0.77$, NCM index). A similar outcome was also obtained when using the sonorant/obstruent segmentation. Sonorant segments contain vowels and a small portion of consonants (e.g., semivowels and nasals), with vowels comprising of about 79% (in duration) of the sonorant segments. Better intelligibility prediction was obtained with sonorant segments ($r = 0.84$ with NCM, and $r = 0.85$ with CSII) than with obstruent segments ($r = 0.64$ with NCM, and $r = 0.69$ with CSII). The difference in correlation was statistically significant (see Tables II and III). To some extent, these outcomes are consistent with the findings of Kewley-Port *et al.* (2007) regarding the contributions of vowels vs consonants to sentence intelligibility. It should be pointed out, however, that the vowel/consonant boundary assignments and the corpora used were different. Furthermore, the RMS-based segmentation did not yield

TABLE V. Duration distribution (in percentage) of vowels and consonants across the three types of RMS-level segments. The nonphoneme segments denote the pauses and epenthetic silence contained in the TIMIT sentences.

				Consonants				
	Vowels	Consonants	Non-phoneme	Semivowels	Nasals	Stops	Fricatives	Affricates
H-level	79.5%	20.5%	0.0%	14.9%	0.5%	1.2%	3.3%	0.7%
M-level	46.2%	53.7%	0.1%	12.7%	9.4%	9.3%	19.4%	2.8%
L-level	6.8%	84.5%	8.7%	5.8%	11.3%	33.3%	31.6%	2.6%

TABLE VII. The average duration (in percentage) for RMS-level and sonorant/obstruent based segmentation for IEEE sentences.

	RMS-level			SON/OBS	
	H-level	M-level	L-level	Sonorants	Obstruents
Average duration	29.1%	28.9%	39.1%	57.2%	36.5%

precise vowel/consonant boundaries as available in the TIMIT corpus as it is not based on *a priori* acoustics/phonetics knowledge. Overall, the above data suggest that better sentence intelligibility prediction can be obtained when using vowel-dominated segments. In fact, intelligibility prediction is better compared to the control condition ($r = 0.85$ vs $r = 0.82$, CSII measure) in which all segments are used. Although high correlation was obtained when including vowel-dominated segments in the intelligibility measures, that was not the highest correlation attained in this study suggesting that other segmentations (not based on vowels or consonants) provided better intelligibility prediction, and this is discussed next.

B. Contributions of vowel-consonant boundaries to sentence intelligibility prediction

Relatively lower correlations were obtained with both NCM and CSII measures when using L-level and H-level RMS segments than when using M-level RMS segments (i.e., $r = 0.89$ and 0.91 for the NCM and CSII measures, respectively). As shown in Table V, neither vowels nor consonants dominated the M-level segments as equal duration of vowels and consonants were present. This raises the question: What contributed to the improvement in sentence intelligibility prediction with M-level segmentation? As shown in Table VI, the M-level segments contained a larger number of vowel-consonant boundaries compared to those present in the H- and L-segments (see Fig. 1). The H-level segments, for instance, only contained 18.9% of C-V boundaries suggesting that these segments captured the inner (and perhaps steady-state) portion of the vowels. Hence, based on the relatively higher occurrence of C-V boundaries in M-level segments (i.e., Table VI), we hypothesized that the vowel-consonant boundaries accounted for much of the variance in intelligibility scores and contributed to the higher correlation.

The highest correlations with both NCM and CSII measures were obtained when including segments containing a large number of consonant-vowel boundaries (see Table VI). To some extent, this outcome is consistent with that of Stilp and Kluender (2010) since these segments contain a great deal of spectral change, i.e., they have the highest entropy. Compared to the scaled-entropy metric, however, which captures spectral change present within *and* between vowels and consonants, the M-level segmentation focuses primarily on consonant-vowel transitions. In Fig. 1 for instance, an H-entropy segment was identified within the vowel (see $t = 0.6$ to 0.7 s) capturing the F2 and F3 formant movements in /eh r/. This same vocalic segment, however, was classified as an H-level segment and not as an M-level segment.

Lee and Kewley-Port (2009) recently examined the intelligibility of interrupted sentences (using the noise-replacement paradigm) preserving four different types of sub-segmental cues, namely steady-state cues at centers or transitions of vowel-consonant margins, and vowel onset or offset transitions. They found that intelligibility scores were not significantly different among the various types of sub-segmental information used. Dynamic transition cues did not provide more benefit than quasi-steady-state cues. To some extent, this differs with the finding of the present study on the contribution of C-V boundaries for intelligibility prediction. The duration of sub-segmental cues in Lee and Kewley-Port (2009), however, was constrained to be 50% or 70% of sentence duration. No such duration constraint on the M-level segmentation was imposed in this study. We believe this might account for the discrepancy on the role of C-V boundaries for intelligibility prediction between the present study and the study by Lee and Kewley-Port (2009). Further studies are warranted to assess the effect of segment duration (containing C-V boundaries) to intelligibility prediction.

Both the entropy and M-level methods capture spectral change, however, with the following two differences. First, the M-level segmentation is instantaneous (in that it requires no past temporal or spectral information) and perhaps more accurate (at identifying vowel-consonant boundaries), whereas the entropy-based segmentation requires an 80 ms accumulation of spectra changes prior to the computation of the entropy (Stilp and Kluender, 2010). In Fig. 1 for instance, the computed scaled-entropy was low at $t = 1.6$–1.8 s (segments /v/ /r/ /ix/ /dh/) despite the spectral changes present (i.e., the distinct /v/ to /r/ transition). Entropy was found to be low because it was computed based on spectral information accumulated in the past 80 ms, which in this case happened to contain stationary (steady) vowel information (vowel /aw/ followed by vowel /ow/). Put differently, the scaled entropy computation seems to be influenced by past contextual information particularly when the immediate past spectral information is relatively stationary and subsequently of low entropy. In contrast, the M-level segmentation method is able to capture the rapid spectral changes, such as those occurring at $t = 1.6$–1.8 s (segments /v/ /r/ /ix/ /dh/) in Fig. 1. The second difference is that the entropy-based segmentation is designed to capture all spectral changes occurring within and between vowels and consonants, whereas the M-level segmentation captures primarily spectral changes occurring at the consonant-vowel boundaries. To some extent, these spectral changes are the most distinct (and largest) and perhaps contribute the most to speech recognition in noise. Furthermore, these spectral changes are perhaps more robust in the presence of noise compared to the changes occurring within vocalic segments, which might be masked by noise.

In terms of mere number of CV transitions, both M-level and H-entropy segmentations yielded the same number (see Table VI). Both methods captured about 50% of the CV transitions, but the characteristics of these transitions were not the same (see example in Fig. 1). The overall duration of the segments falling in those transitions was longer for the H-entropy method than the M-level method (see Table VIII). An average 85% of the total duration of the H-entropy segments was classified as CV transitions compared to 55% of

TABLE VIII. Comparison of duration distribution (in percentage) of CV transitions between M-level and H-entropy segmentation methods. The H-entropy was computed with proportion coefficient $p = 60\%$.

		Non-CV-transition		
	CV-transition	Vowels	Consonants	Non-phoneme
M-level	55.0%	15.9%	29.1%	0%
H-entropy	84.8%	6.9%	7.8%	0.5%

the total duration of the M-level segments. This was expected given that the H-entropy method captures all spectral changes including transitions within vowels, while the M-level method captures only transitions from vowels to consonants and vice versa. We believe that it was this difference that contributed to the difference in correlations between the H-entropy and M-level segmentation methods when incorporated in the computation of the intelligibility measures (i.e., NCM and CSII) (Table I). A higher correlation was obtained with the M-level method ($r = 0.89$ and 0.91 for NCM and CSII, respectively) than the H-entropy method ($r = 0.82$ for both NCM and CSII). A second potential reason for the difference in correlations is the variability in the spectra/temporal characteristics of the segments included in the computation of the measures of the two methods. A higher variability (i.e., large proportion of non-homogeneous segments) was present in the H-entropy method and a lower variability (i.e., relatively large proportion of homogeneous segments) was present in the M-level method. This difference in variability might have affected the predictive power of the intelligibility measures.

Despite their differences, the entropy and RMS segmentations share something in common: both methods do away with the traditional phonetic distinction between vowels and consonants. The RMS segmentation makes no use of *a priori* knowledge about the temporal or spectral characteristics of vowels or consonants other than the fact that vowels are generally higher in intensity than consonants. In contrast, the sonorant/obstruent segmentation assumes knowledge of the presence of periodicity (F0) as introduced by the vibration of the vocal folds in voiced sounds or aperiodicity in unvoiced sounds. It is worth noting that while the present study showed the importance of vowel-consonant transitions as captured by M-level segmentation in English, the contribution (and importance) of vowel-consonant transitions in intelligibility prediction might shift in other languages. In tonal languages (e.g., Mandarin Chinese), for instance, we have shown previously (Chen and Loizou, 2011b) that the H-level segments carry perceptually more important information and are better predictors of intelligibility than the M-level segments. This was attributed to the increased importance of F0 information needed for reliable tone recognition in Mandarin Chinese. These F0 cues are present in the vowel-dominated H-level segments.

In brief, the present data suggest that the consonant-vowel boundaries contributed the most in terms of (English) sentence intelligibility prediction in noise. The importance of these boundaries as potential acoustic landmarks in the signal has also been implicated in lexical access models (Stevens, 2002), at least in quiet. Stevens' lexical access model (Stevens, 2002) comprises of multiple steps with the first step being responsible for signal segmentation, i.e., the signal is segmented into acoustic landmarks (present at the boundaries of the vowels, consonants, and glide segments) based on detection of peaks and spectral discontinuities in the signal. This initial step is important because if the acoustic landmarks are perceptually not clear or distinct owing to corruption of the signal by external noise, it would affect the subsequent stages of the model. In noise and/or reverberation, for instance, these landmarks might be blurred and arguably could play an even more important role (Li and Loizou, 2008) in lexical segmentation than in quiet. Li and Loizou (2008) assessed the contribution of acoustic landmarks in noise, and in particular those signified by spectral discontinuities at the onsets/offsets of obstruent consonants, and found that listeners received a large benefit in intelligibility when provided with access to these landmarks in otherwise noise-masked sentences.

Overall, the data from the present study suggest that intelligibility models need to account, and perhaps place more emphasis, on spectral changes that can be quantified using either the M-level or scaled-entropy segmentation methods. The main difference between the two methods is the emphasis placed on the various forms of spectral change occurring throughout the utterance. While the scaled-entropy method makes no distinction between spectral changes occurring at vowels/consonants or at vowel/consonant transitions, the M-level method places more emphasis on the vowel/consonant transitions wherein the spectral change is often most prominent, and perhaps most robust in the presence of noise. Data collected in our study suggests that the M-level method better predicts the listeners' intelligibility scores, and the difference in correlation was found to be statistically significant (Tables II and III).

ACKNOWLEDGMENTS

This research was supported by Grant No. R01 DC010494 from the National Institute of Deafness and other Communication Disorders, NIH. The authors are grateful to the two reviewers who provided valuable feedback that significantly improved the presentation of the manuscript.

[1]The STI measure (Steeneken and Houtgast, 1980) is based on the idea that the reduction in intelligibility caused by additive noise or reverberation can be modeled in terms of the reduction in temporal envelope modulations. The STI metric has been shown to predict successfully the effects of reverberation, room acoustics, and additive noise (e.g., Steeneken and Houtgast, 1982; Houtgast and Steeneken, 1985). In its original form (Houtgast and Steeneken, 1971), the STI measure used artificial signals (e.g., sinewave-modulated signals) as probe signals to assess the reduction in signal modulation in a number of frequency bands and for a range of modulation frequencies (0.6–12.5 Hz) known to be important for speech intelligibility.

[2]In the original form of the CSE calculation (Stilp and Kluender, 2010), 16 ms slices of TIMIT sentences (sampled at 16 kHz) were passed through 33 filters, spaced one ERB apart, spanning 26 to 7743 Hz. This bandwidth is twice the bandwidth used in the present study (300–3400 Hz) and the number of ro-ex filters used in the present study is half as many as that used by Stilp and Kluender (2010). In the CSE calculation, however,

high-frequency information is attenuated relative to the emphasized low-frequency information. Hence, there exists the possibility that both analyses yield qualitatively similar results, and further work is needed to confirm this.

ANSI. (**1997**). "Methods for calculation of the speech intelligibility index," S3.5–1997 (American National Standards Institute, New York).

Arehart, K., Kates, J., Anderson, M., and Harvey, L. (**2007**). "Effects of noise and distortion on speech quality judgments in normal-hearing and hearing-impaired listeners," J. Acoust. Soc. Am. **122**, 1150–1164.

Chen, F., and Loizou, P. (**2011a**). "Predicting the intelligibility of vocoded speech," Ear Hear. **32**, 331–338.

Chen, F., and Loizou, P. (**2011b**). "Predicting the intelligibility of vocoded and wideband Mandarin Chinese," J. Acoust. Soc. Am. **129**, 3281–3290.

Cole, R., Yan, Y., Mak, B., Fanty, M., and Bailey, T. (**1996**). "The contribution of consonants versus vowels to word recognition in fluent speech," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 853–856.

Fogerty, D., and Kewley-Port, D. (**2009**). "Perceptual contributions of the consonant-vowel boundary to sentence intelligibility," J. Acoust. Soc. Am. **126**, 847–857.

Garofolo, J., Lamel, L., Fisher, W., Fiscus, J. G., Pallett, D. S., Dahlgren, N. L., and Zue, V. (**1993**). *TIMIT Acoustic-Phonetic Continuous Speech Corpus* (Linguistic Data Consortium, Philadelphia, PA).

Goldsworthy, R., and Greenberg, J. (**2004**). "Analysis of speech-based speech transmission index methods with implications for nonlinear operations," J. Acoust. Soc. Am. **116**, 3679–3689.

Hirsch, H., and Pearce, D. (**2000**). "The aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions," in *ISCA Tutorial and Research Workshop ASR2000*, Paris, France.

Houtgast, T., and Steeneken, H. J. M. (**1971**). "Evaluation of speech transmission channels by using artificial signals," Acustica **25**, 355–367.

Houtgast, T., and Steeneken, H. J. M. (**1985**). "A review of the MTF concept in room acoustics and its use for estimating speech intelligibility in auditoria," J. Acoust. Soc. Am. **77**, 1069–1077.

Hu, Y., and Loizou, P. (**2007**). "A comparative intelligibility study of single-microphone noise reduction algorithms," J. Acoust. Soc. Am. **122**, 1777–1786.

IEEE. (**1969**). "IEEE recommended practice for speech quality measurements," IEEE Trans. Audio Electroacoust. **17**, 225–246.

Kates, J. (**1992**). "On using coherence to measure distortion in hearing aids," J. Acoust. Soc. Am. **91**, 2236–2244.

Kates, J., and Arehart, K. (**2005**). "Coherence and the speech intelligibility index," J. Acoust. Soc. Am. **117**, 2224–2237.

Kewley-Port, D., Burkle, T. Z., and Lee, J. H. (**2007**). "Contribution of consonant versus vowel information to sentence intelligibility for young normal-hearing and elderly hearing-impaired listeners," J. Acoust. Soc. Am. **122**, 2365–2375.

Kryter, K. D. (**1962a**). "Methods for the calculation and use of the articulation index," J. Acoust. Soc. Am. **34**, 1689–1697.

Kryter, K. D. (**1962b**). "Validation of the articulation index," J. Acoust. Soc. Am. **34**, 1698–1706.

Lee, J. H., and Kewley-Port, D. (**2009**). "Intelligibility of interrupted sentences at subsegmental levels in young normal-hearing and elderly hearing-impaired listeners," J. Acoust. Soc. Am. **125**, 1153–1163.

Li, N., and Loizou, P. (**2008**). "The contribution of obstruent consonants and acoustic landmarks to speech recognition in noise," J. Acoust. Soc. Am. **124**, 3947–3958.

Loizou, P. (**2007**). *Speech Enhancement: Theory and Practice* (CRC, Boca Raton, FL), pp. 589–592.

Ma, J., Hu, Y., and Loizou, P. (**2009**). "Objective measures for predicting speech intelligibility in noisy conditions based on new band-importance functions," J. Acoust. Soc. Am. **125**, 3387–3405.

Maddieson, I. (**1984**). *Patterns of Sounds* (Cambridge University Press, New York), Chap. 1, pp. 1–24.

Patterson, R. D., Nimmo-Smith, I., Weber, D. L., Milroy, R. (**1982**). "The deterioration of hearing with age: Frequency selectivity, the critical ratio, the audiogram, and speech threshold," J. Acoust. Soc. Am. **72**, 1788–1803.

Pavlovic, C. V. (**1987**). "Derivation of primary parameters and procedures for use in speech intelligibility predictions," J. Acoust. Soc. Am. **82**, 413–422.

Rhebergen, K. S., and Versfeld, N. J. (**2005**). "A speech intelligibility index-based approach to predict the speech reception threshold for sentences in fluctuating noise for normal-hearing listeners," J. Acoust. Soc. Am. **117**, 2181–2192.

Steeneken, H., and Houtgast, T. (**1980**). "A physical method for measuring speech transmission quality," J. Acoust. Soc. Am. **67**, 318–326.

Steeneken, H., and Houtgast, T. (**1982**). "Some applications of the speech transmission index (STI) in auditoria," Acustica **51**, 229–234.

Steiger, J. H. (**1980**). "Tests for comparing elements of a correlation matrix," Psychol. Bull. **87**, 245–251.

Stevens, K. (**2002**). "Toward a model for lexical access based on acoustic landmarks and distinctive features," J. Acoust. Soc. Am. **111**, 1872–1891.

Stilp, C. E., and Kluender, K. R. (**2010**). "Cochlea-scaled entropy, not consonants, vowels or time, best predicts speech intelligibility," Proc. Natl. Acad. Sci. U.S.A. **107**, 12387–12392.

Strange, W., Jenkins, J. J., and Johnson, T. L. (**1983**). "Dynamic specification of coarticulated vowels," J. Acoust. Soc. Am. **74**, 695–705.

Taal, C., Hendriks, R., Heusdens, R., and Jensen, J. (**2010**). "A short-time objective intelligibility measure for time-frequency weighted noisy speech," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 4214–4217.